HELP International Humanitarian NGO

Presented By,
Navya Somesh

Problem Statement

- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Business Goal

- Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Analysis Approach

- There are nine attributes that needs be analyzed to determine which countries need the aid.
- PCA will be used to find the linear combination between the attributes and to remove multicollinear data and to find the Principal Components which help in Dimensionality Reduction without dropping any attributes.

Analysis Approach

- Scaling will be performed as it is assumed that PCA requires scaled data.
- Based on these Principal Components Clustering will be performed using both KMeans and Hierarchical Clustering.
- Instead of relying only on Kmeans and Hierarchical its better to consider both and the combined results are analyzed.

Principal Component Analysis

- Statistical procedure that converts the possibly correlated variables to Principal Components such that they are
  - Uncorrelated with each other.
  - Linear Combinations of original variables.
  - Maximum Information will be captured without dropping any variables.

Principal Component Analysis

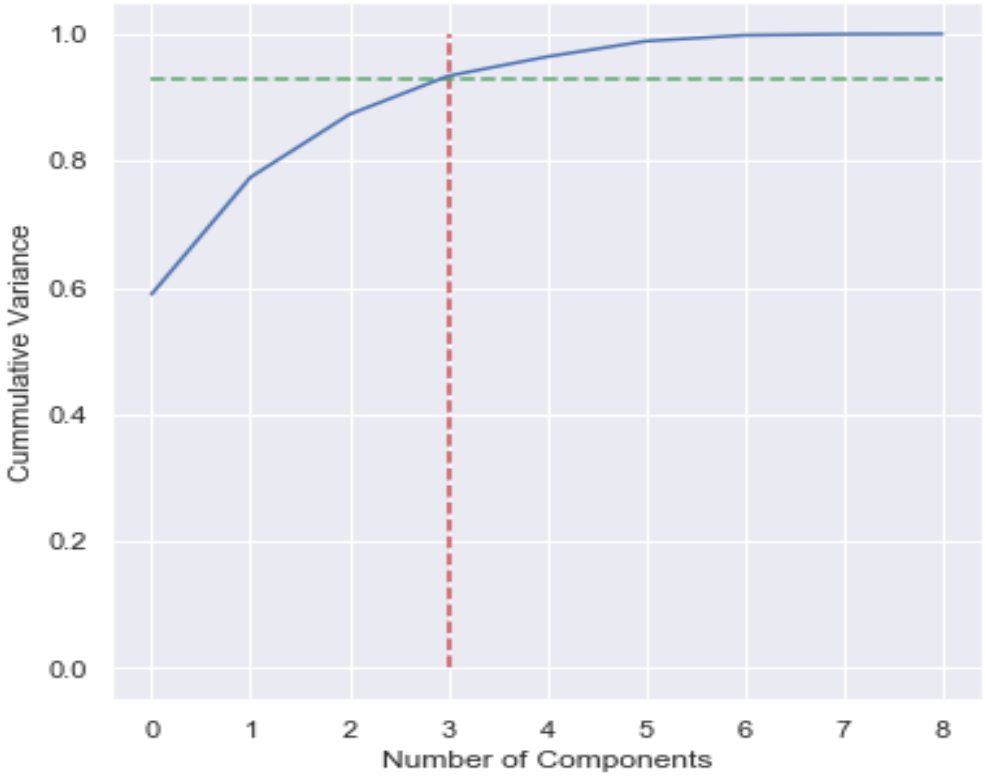
- PCA will be applied on the scaled data and Single Value Decomposition technique will be used.

Scree Plot

- Helps to visualize cumulative variance against the number of components.

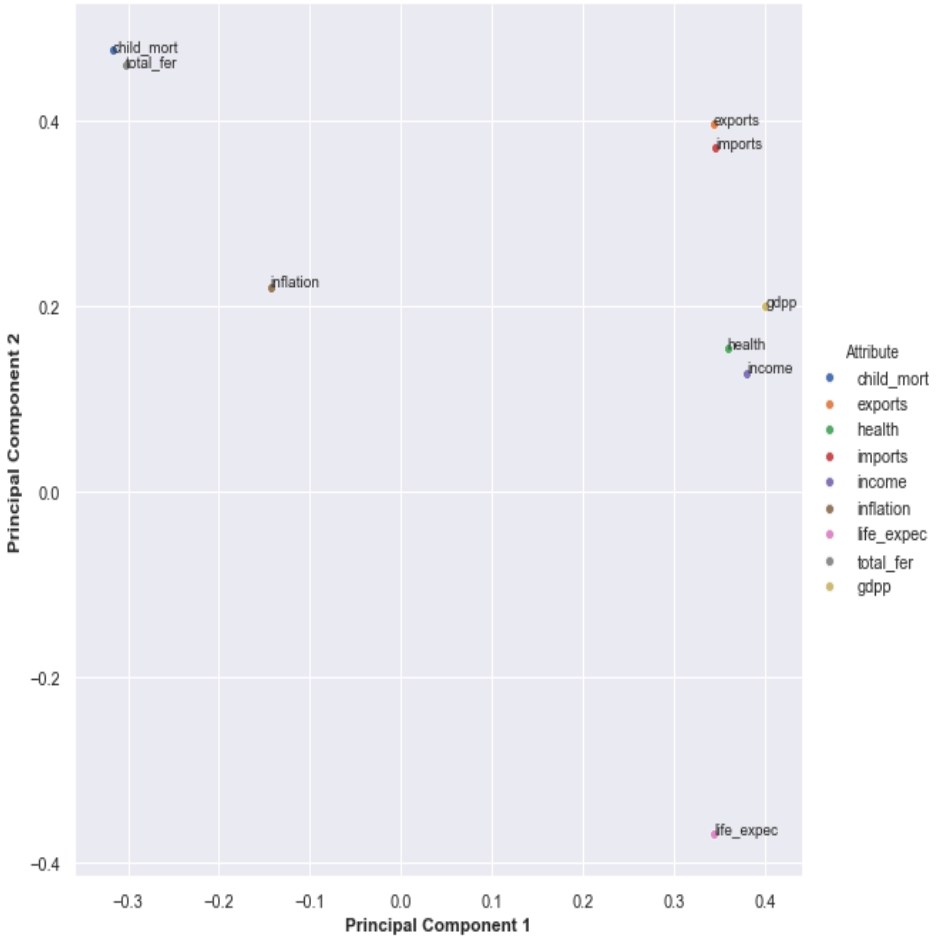
- From the plot it is clear that the Number of Principal Components for analyzing the variables in the dataset will be 3.

- It is evident from the above Scree plot that more than 90% variance is explained by the first 3 Principal Components. Hence, we will use these 3 Principal Components only going forward for Clustering process.



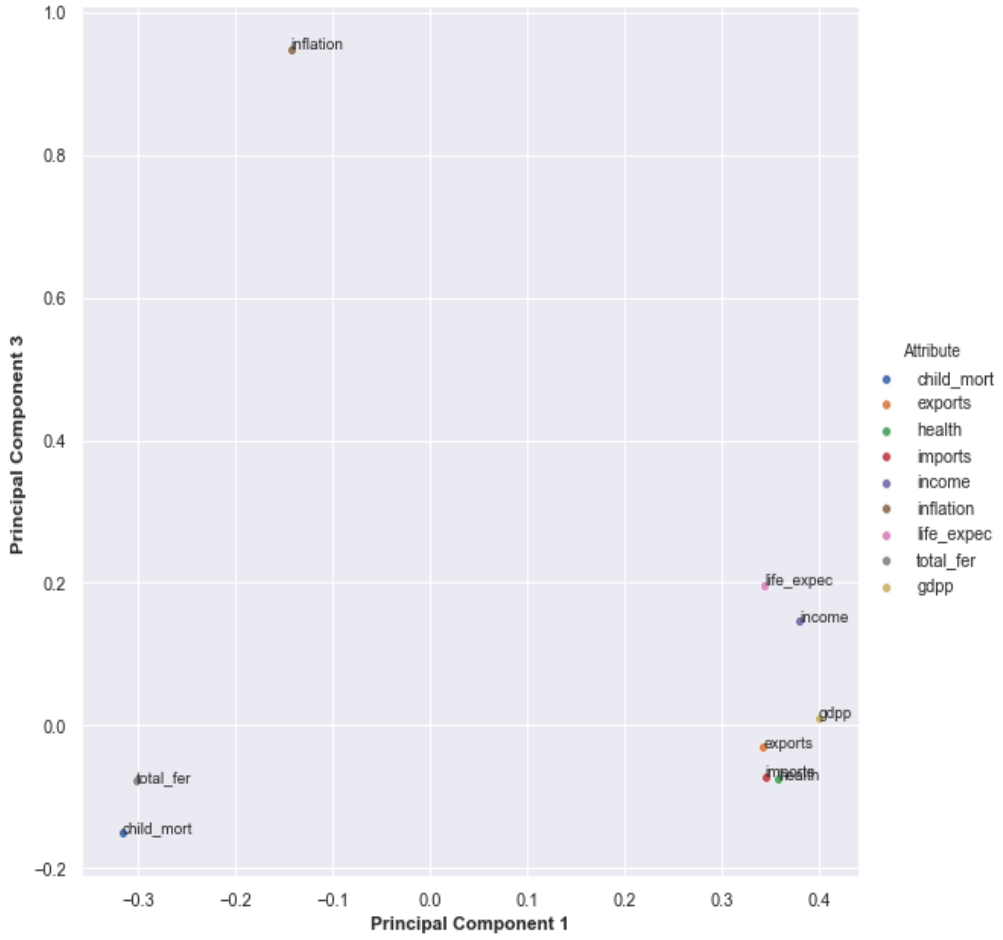
Visualization of PC1 and PC2

- life expectency, income, gdpp and health are very well explained by PC1.
- imports and exports are well explained by both the components PC1 and PC2.
- child mortality and total fertility are well explained by PC2.
- inflation is neither explained by PC1 nor with PC2.



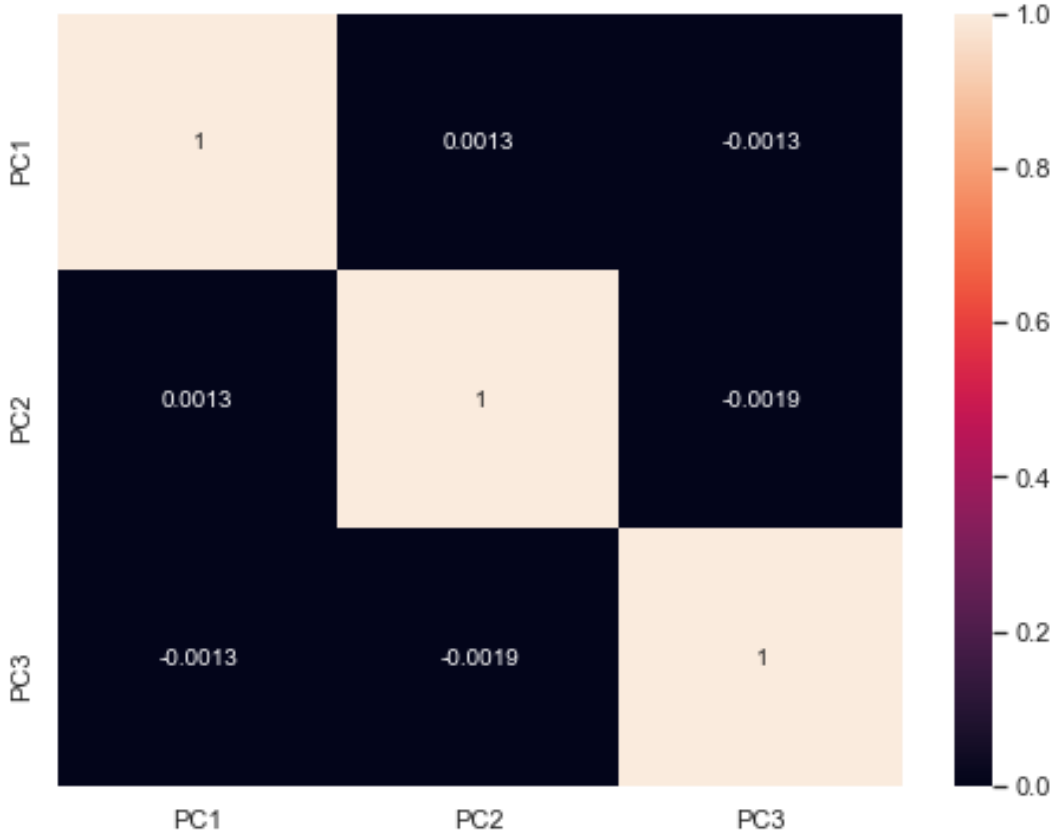
Visualization of PC1 and PC3

- inflation is well explained by PC3.
- Since 90% variance is explained by 3 principal components, dataframe is built by using those 3 components only.



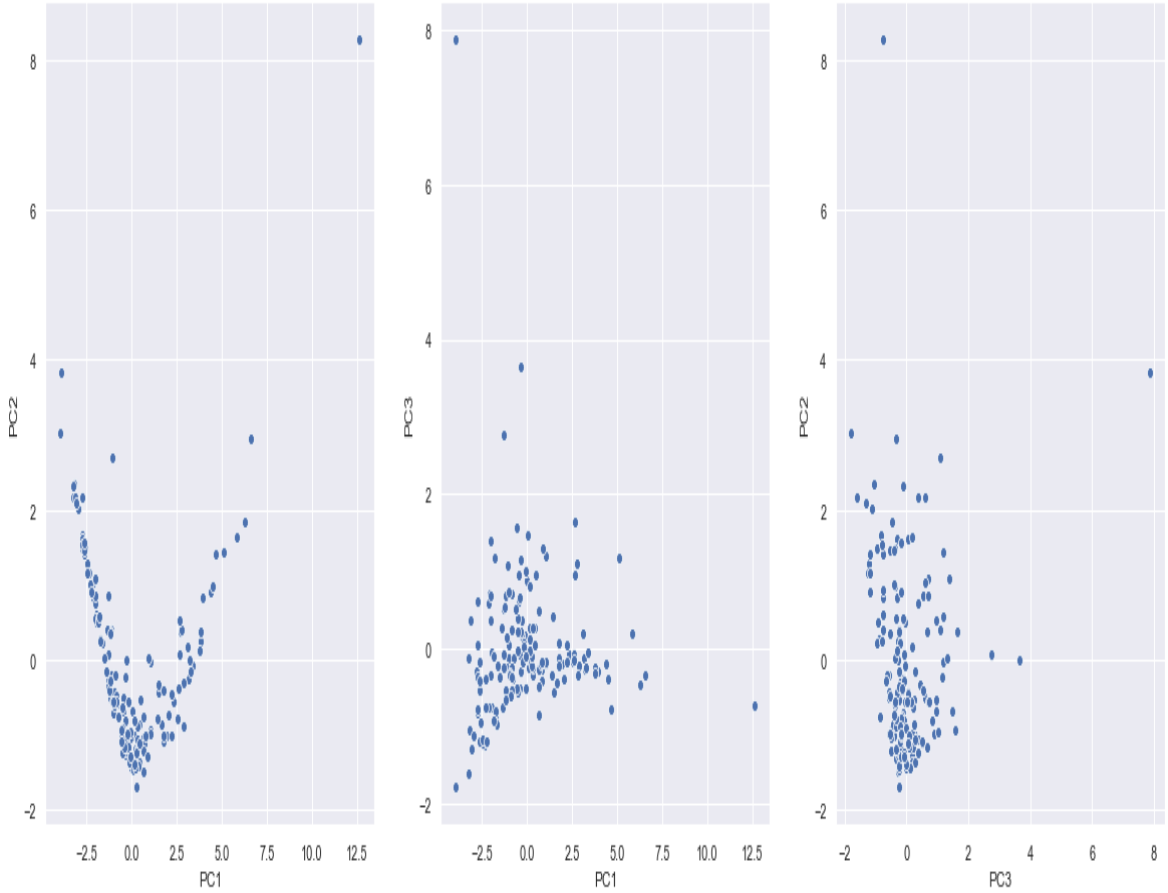
Multicollinearity After PCA

- As we can see from above heatmap that the correlation among the attributes is almost 0.
- Multicollinearity is eliminated.
- Principal Components have created Linear combination from the variables.
- Maximum information is extracted with dropping any column.



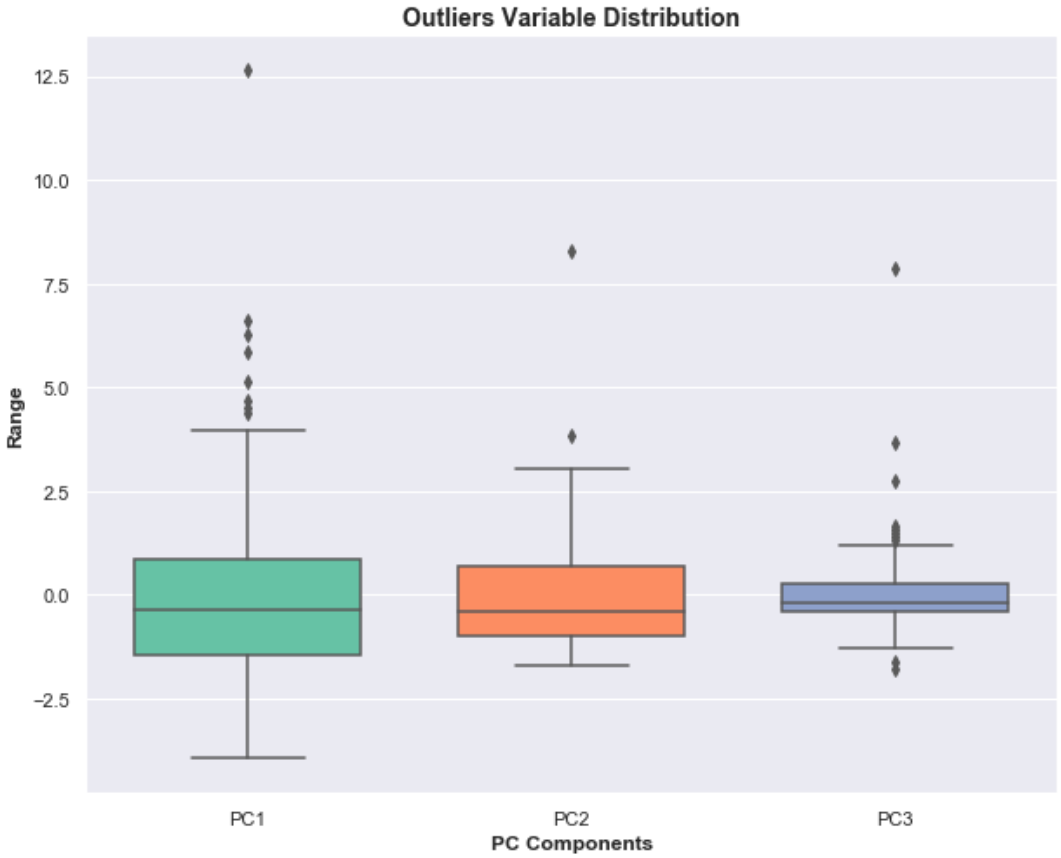
Spread Of Data Across Principal Components

- Inter Cluster distance is more between PC1 and PC2.
- Inter cluster distance is less bewteen PC1,PC3 and PC3,PC2.



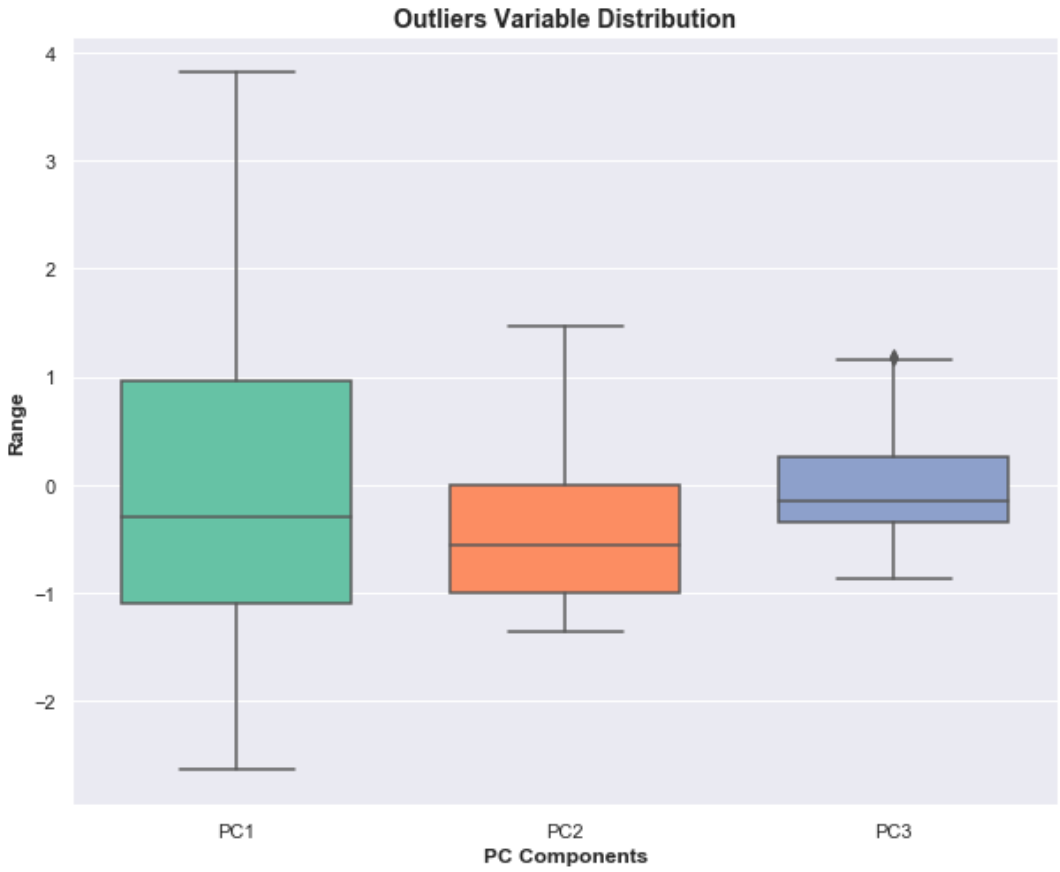
Outlier Analysis

- The outlier treatment results in loss of data but KMeans get affected by outliers and we may not get optimal clustering.
- Hence Outliers will be treated by using IQR method.



Outlier Treatment

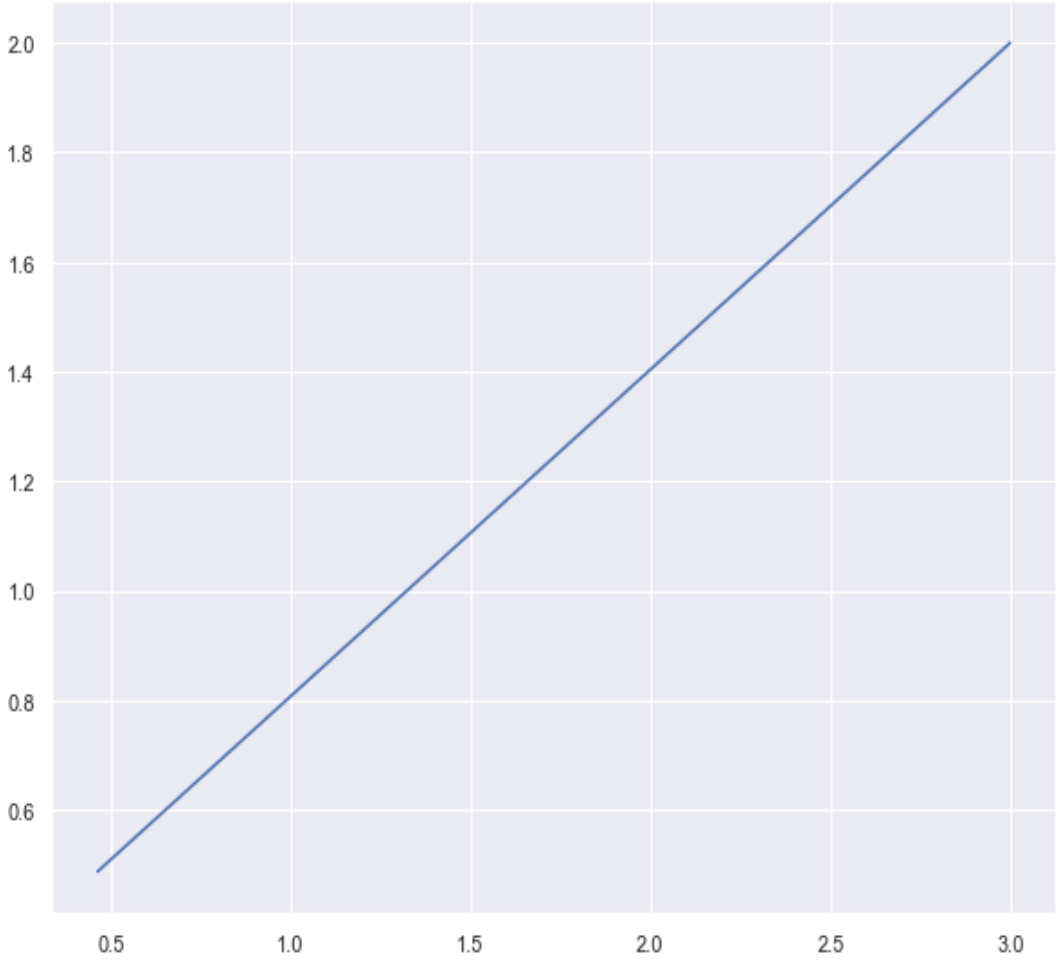
- IQR method is used for outlier treatment.
- The quantile values between 0.05 and 0.95 is considered.
- All the values below 0.05 and above 0.95 are ignored.



K-Means Clustering

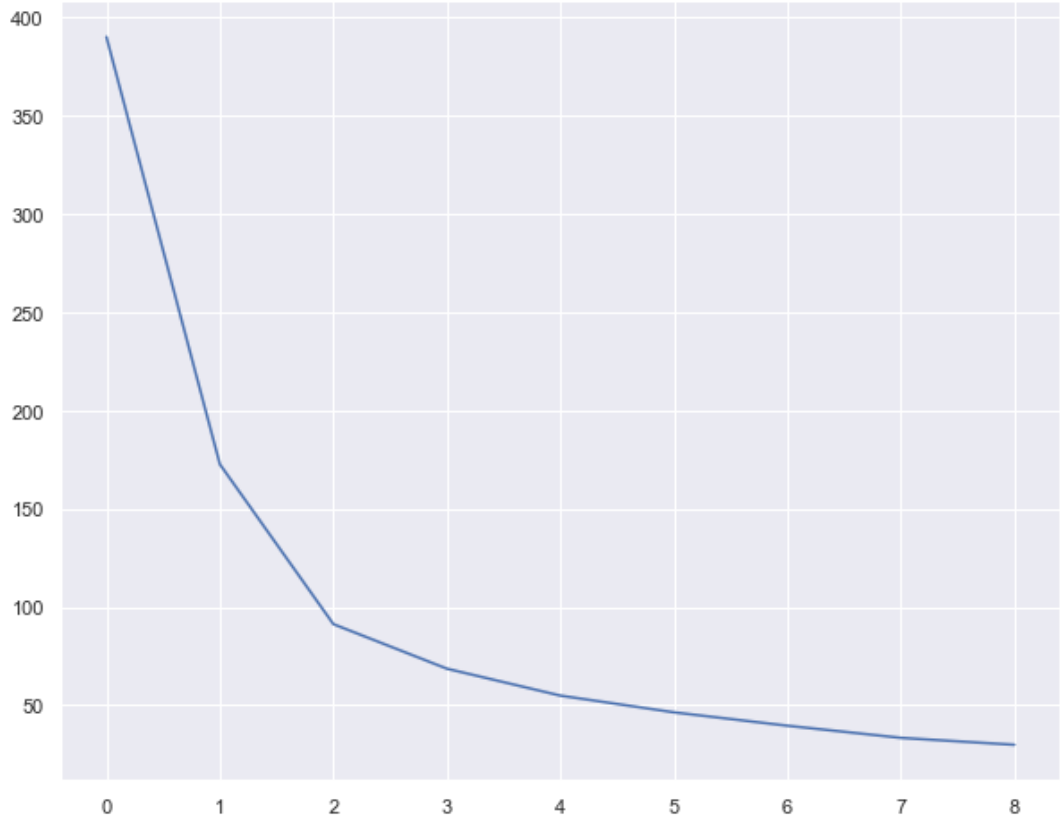
- Clustering refers to collection of data points aggregated together because of certain similarities.
- K-Means groups similar data points together and discover the underlying pattern.
- To determine the OPTIMAL K Value we use
  - Silhouette and
  - Elbow Curve

- Silhouette Score : we take the peak value from the score which defines the maximum/optimal K values of Clustering required.
  - The peak is at 3.
  - The required number of cluster is 3.



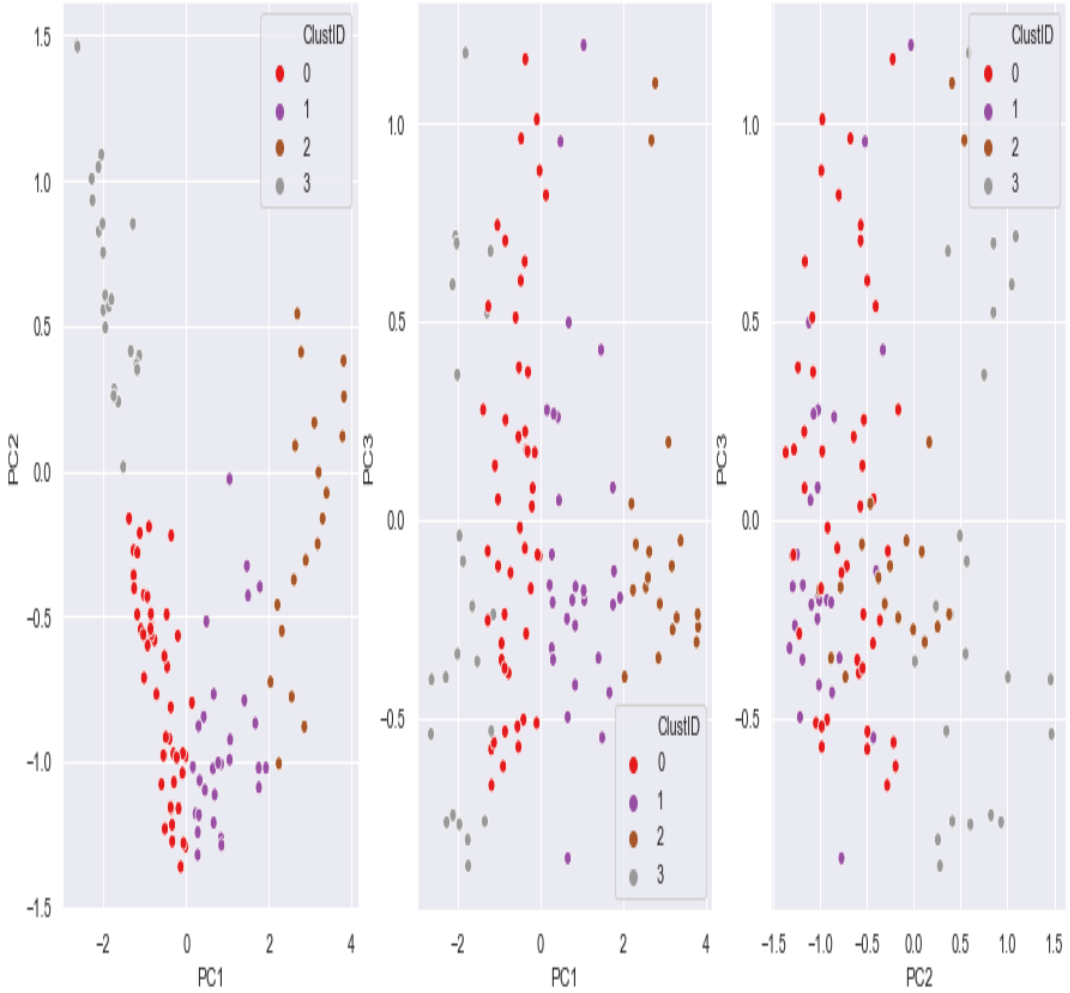
Elbow Curve

- We need to consider the curve value.
- First elbow curve is at 1, cluster 1 doesn't make any sense.
- Second is at 4 so I'm considering k=4
- Looking at the above elbow curve it looks good to proceed with either 4 or 5 clusters.



K-Means with K=4

- In plot 1 there is lots of inter-cluster distance between PC1 and PC2 which is not a good sign.
- In plot 3 the inter-cluster distance is too low between PC2 and PC3 which is also not a good sign.

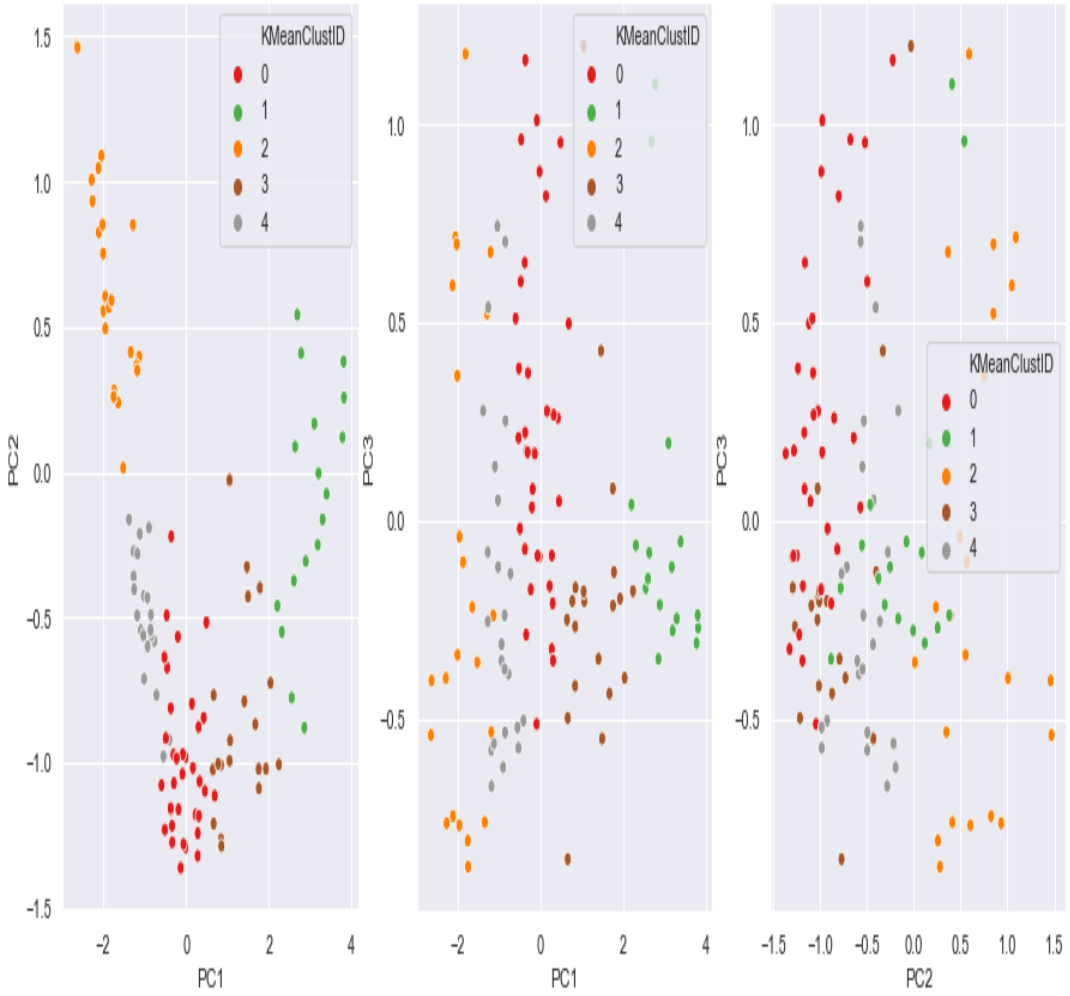


K-Means with K=5

- The clusters formed are easily distinguishable.
- All the drawbacks of k=4 is removed in k=5.
- Also seems like we may get optimal clusters.

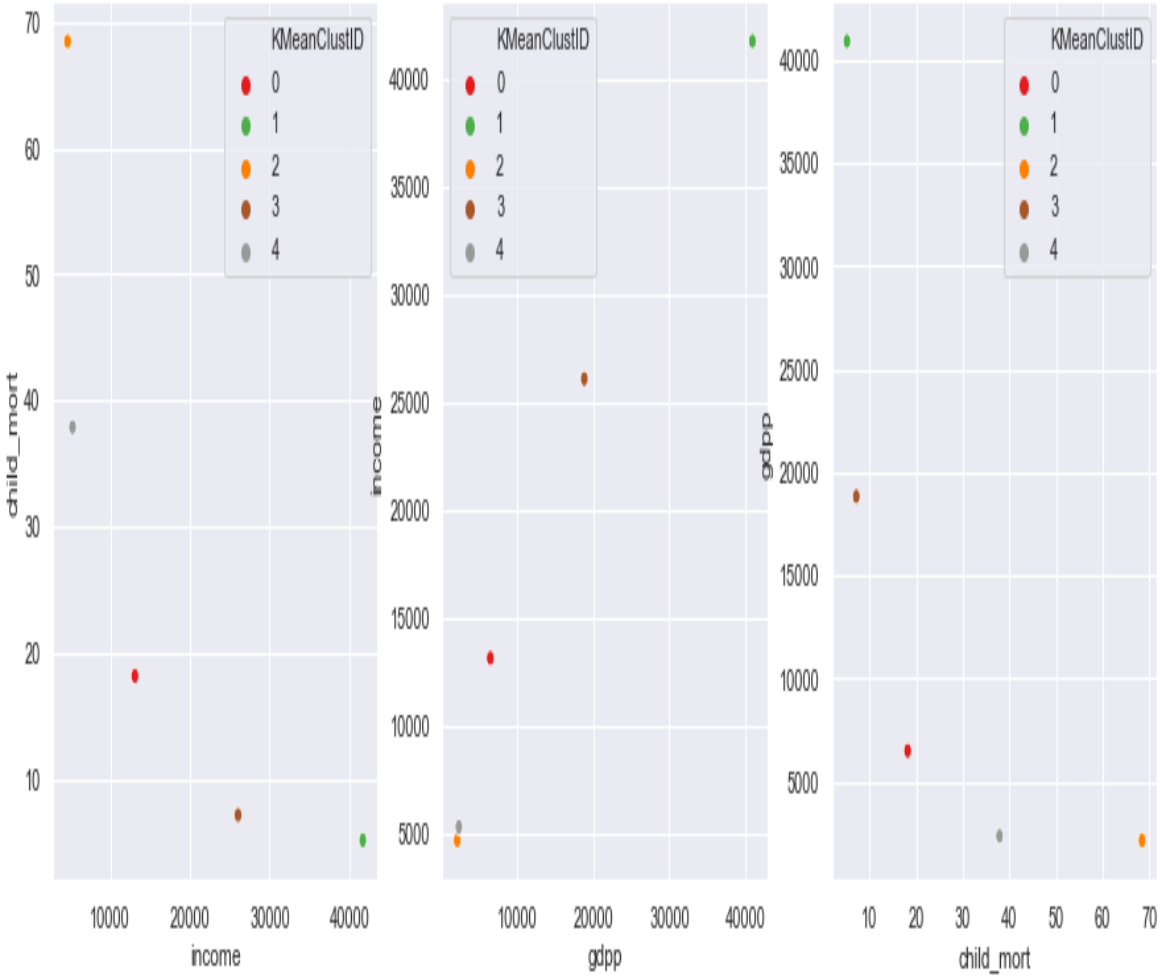
So I'm proceeding with K=5 i.e 5 clusters.

In real time we can get this information of picking up clsters from Clients / Client Manager of how he is focusing to segregate the data



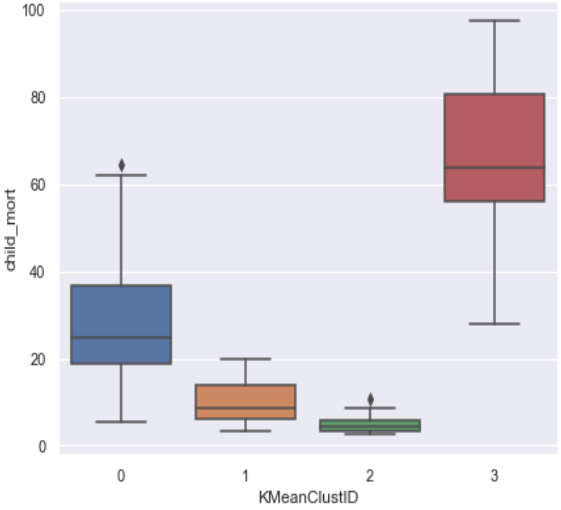
Visualize the spread of Original Data

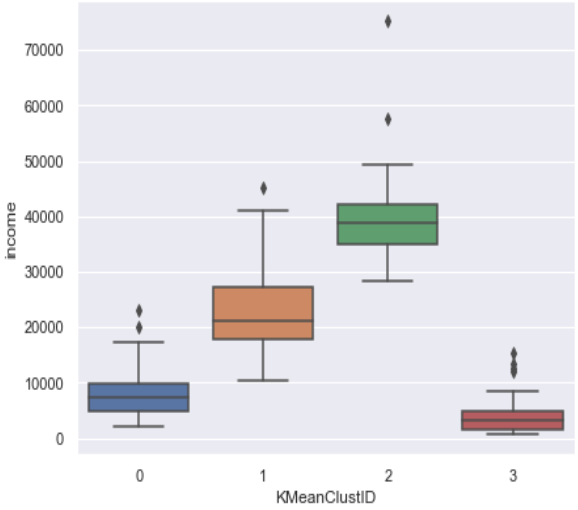
- Cluster 1 has highest income,GDP per capita.
- Cluster 2 has highest child mortality , low income and low gdpp.
- Cluster 3,4 has low income and low gdpp.
- Cluster 0 has low income and low gdpp and low child mortality.

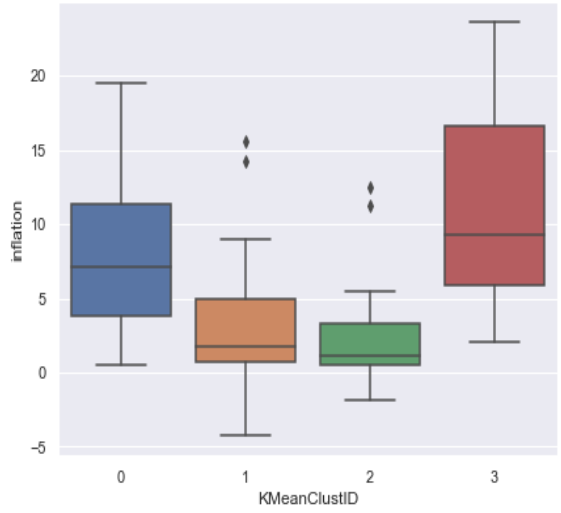


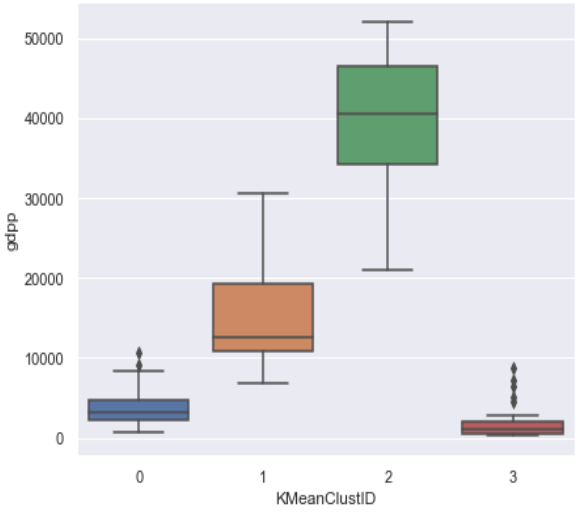
Kmeans attributes to visualize the spread of the data

- Child Mortality is highest for Cluster 0 and Cluster 3.These clusters need some aid.
- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development.
- Income per capita and gdpp seems lowest for countries in clusters 0 and 3. Hence, these countries need some help.



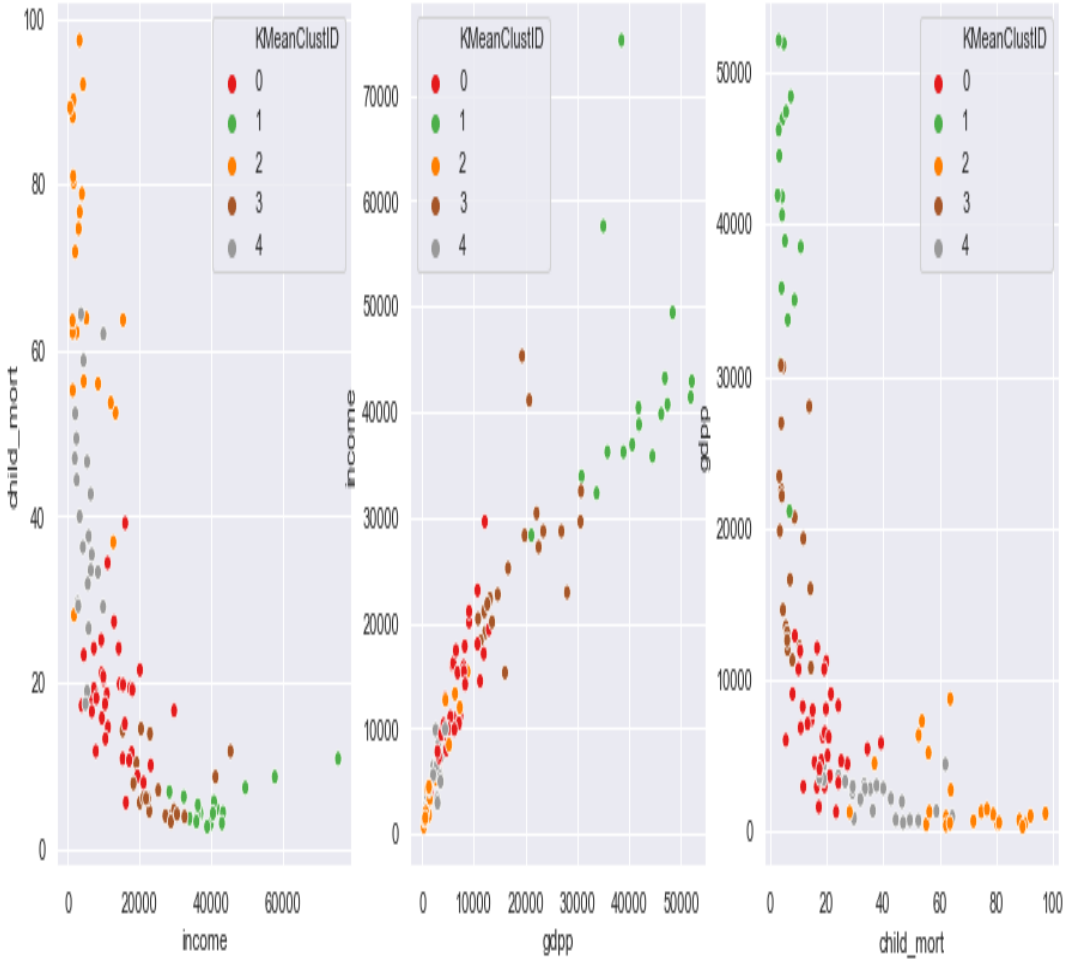






Visualize the mean value of few original attributes

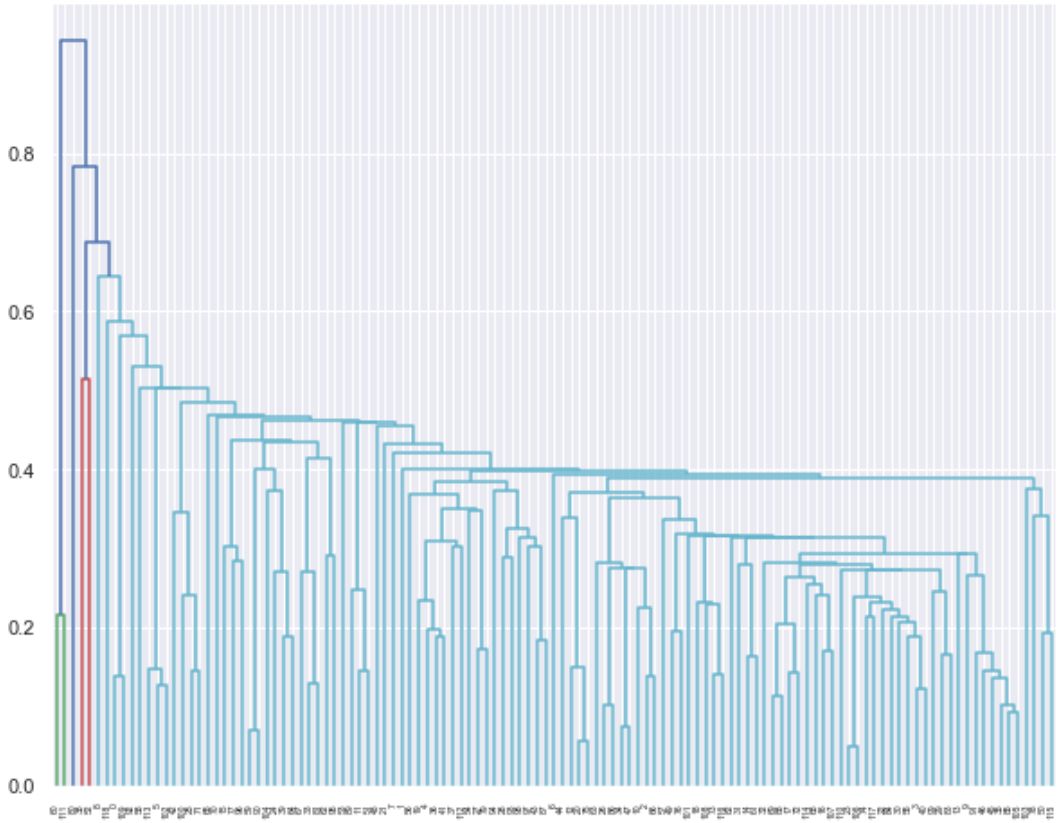
- Child Mortality is highest for Cluster 2 and Cluster 4.These clusters need some aid.
- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development.
- Income per capita and gdpp seems lowest for countries in clusters 2 and 4. Hence, these countries need some help.
- Indicating both results are same. i.e clustering is matching with original data.



Hierarchical Clustering

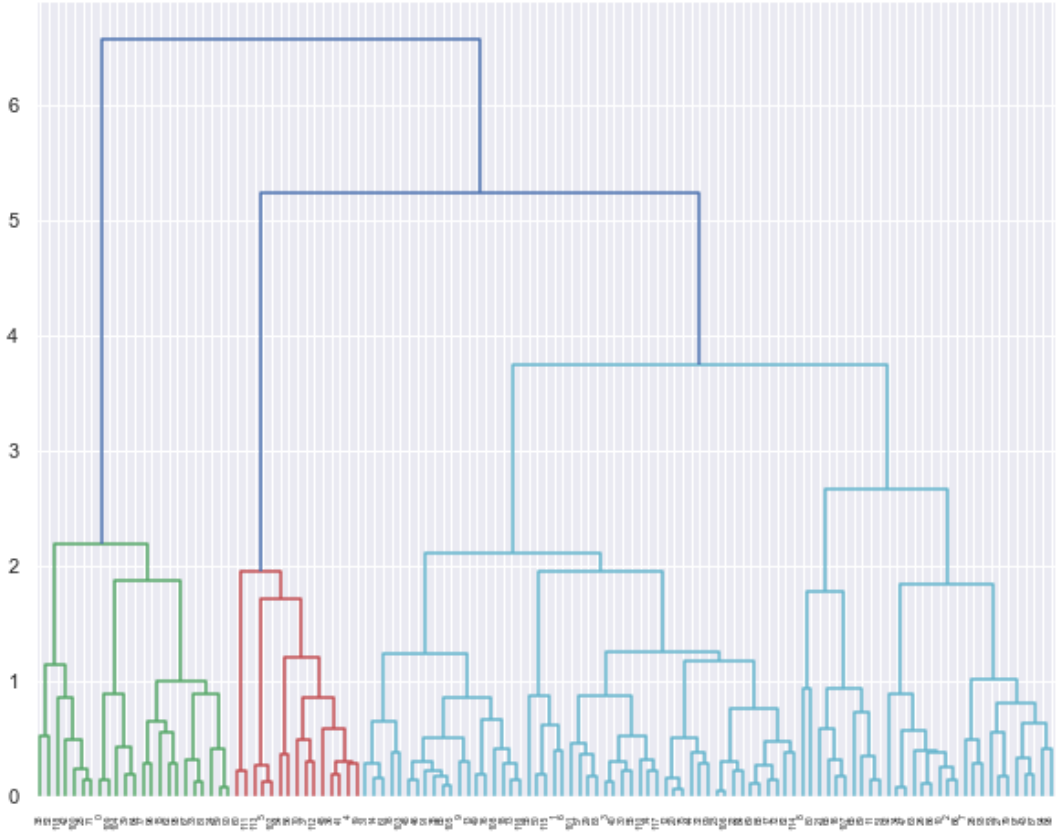
- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.
- There are two types of hierarchical clustering,
- 1. Divisive
- 2. Agglomerative.

- dendogram is not so clear.
- So we shall perform complete linkage to indentify the optimal number of clusters.



Complete Linkage

- dendogram is clear.
- Either the 4 or 5 clusters can be formed.

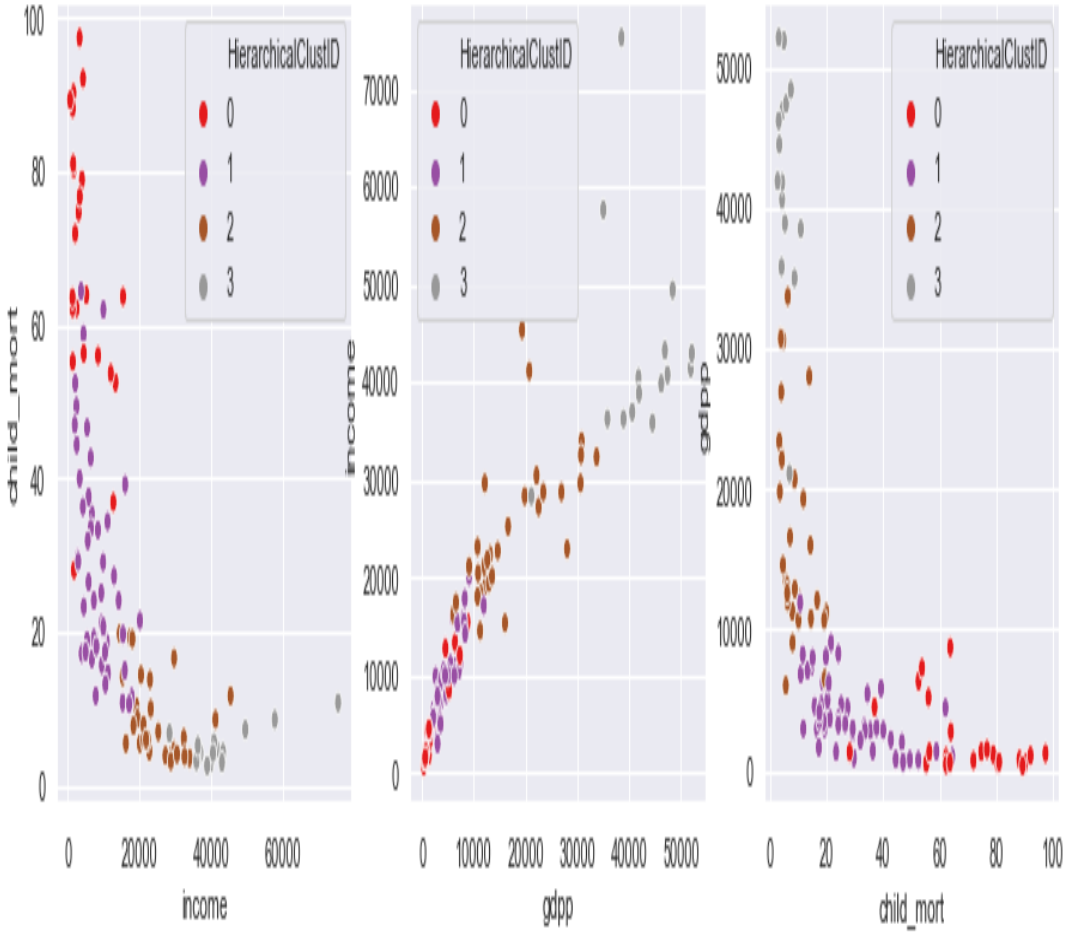


Hierarchical Cluster PC components

- Cluster 4 is not properly formed.
- Seems like majority of countries belong to cluster 0.

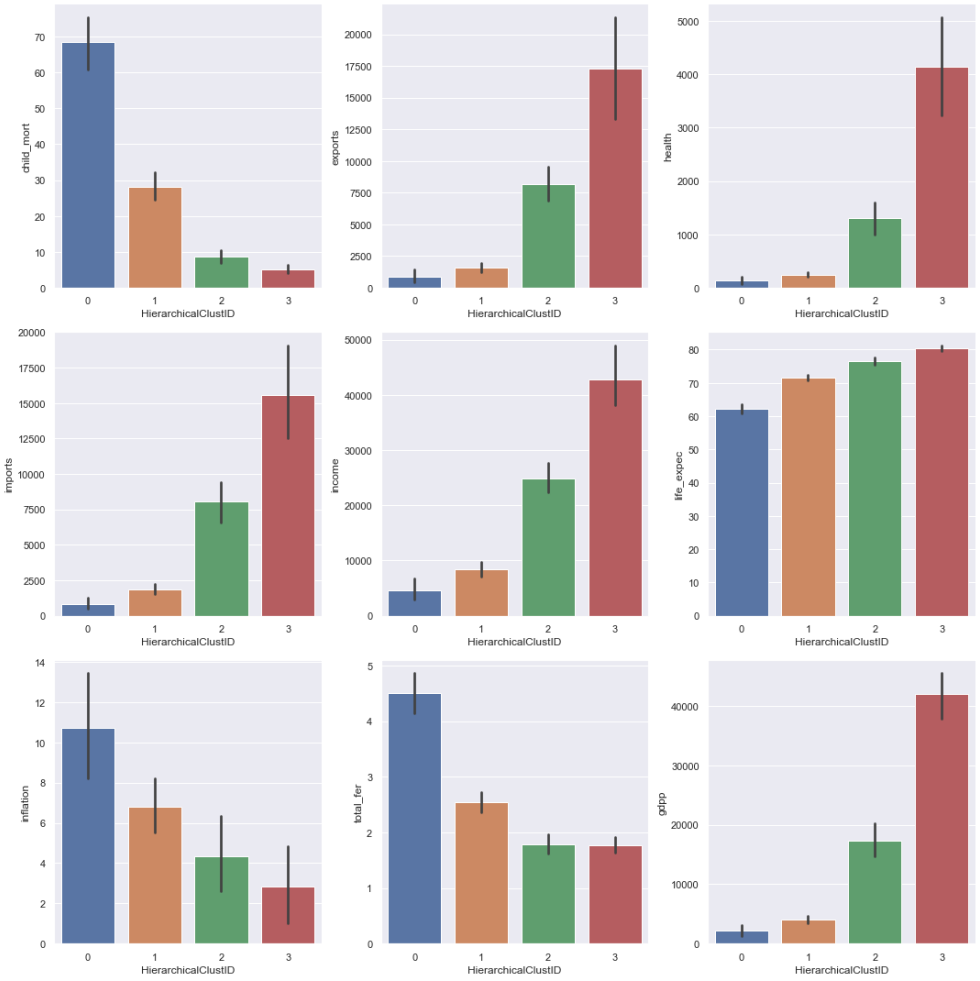
Clusters and Spread of Data

- Cluster 0 countries have higher child mortality and low income , gdpp is also low.
- Cluster 2,3 have low child mortality and high income,gdpp is also relatively high.
- Cluster 1 has high to low wide spread child mortality and low income and low gdpp.



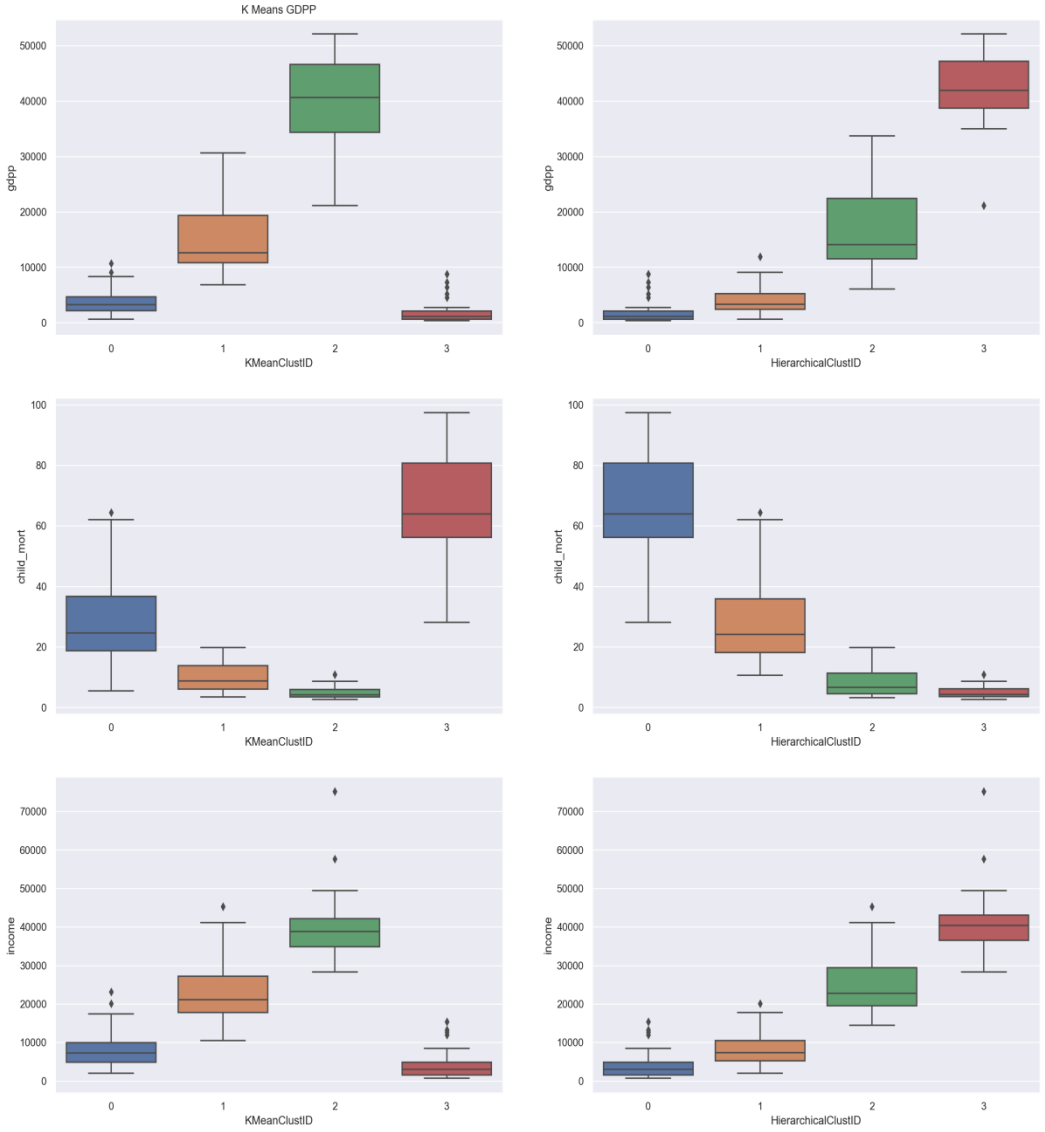
Clusters and Spread of Data Across Variables

- 1. Cluster 0 has high life expentancy and child mortality.
- 2. Cluster 1 has high inflation and fertility.
- 3. Cluster 2 has high GDP per capita,exports and health.
- 4. Cluster 3 doesn't show much variations in any particular aspect.



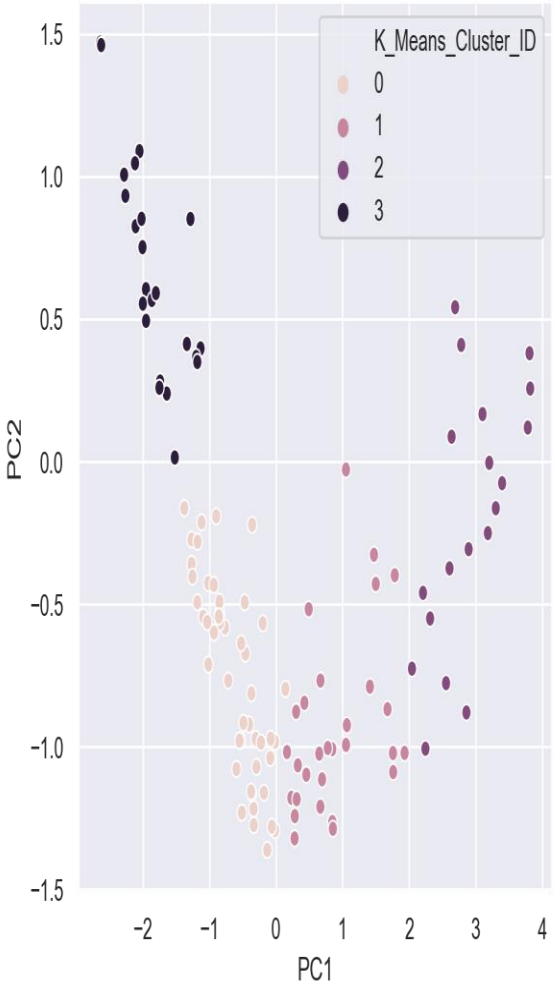
K-Means V/s Hierarchical Clusters

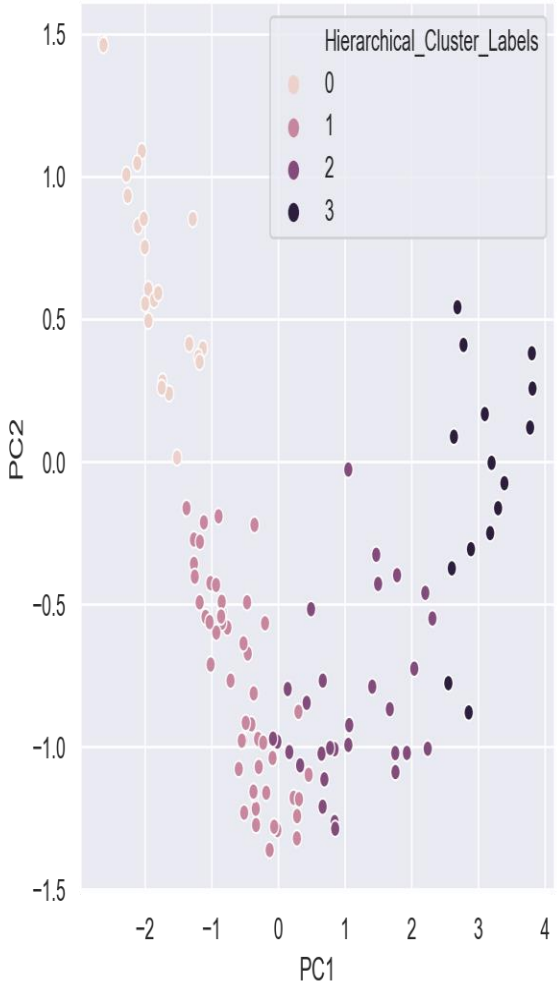
- In K-Means Cluster 0 and 3 are the countries that are in need of aid.
- In Hierarchical Cluster 0 is in dead need of aid compared to other clusters formed.
- The clusters formed in both the cases are not that great but its better in K-means as compared to Hierarchial.



PC of K-Means and Hierarchical Clustering

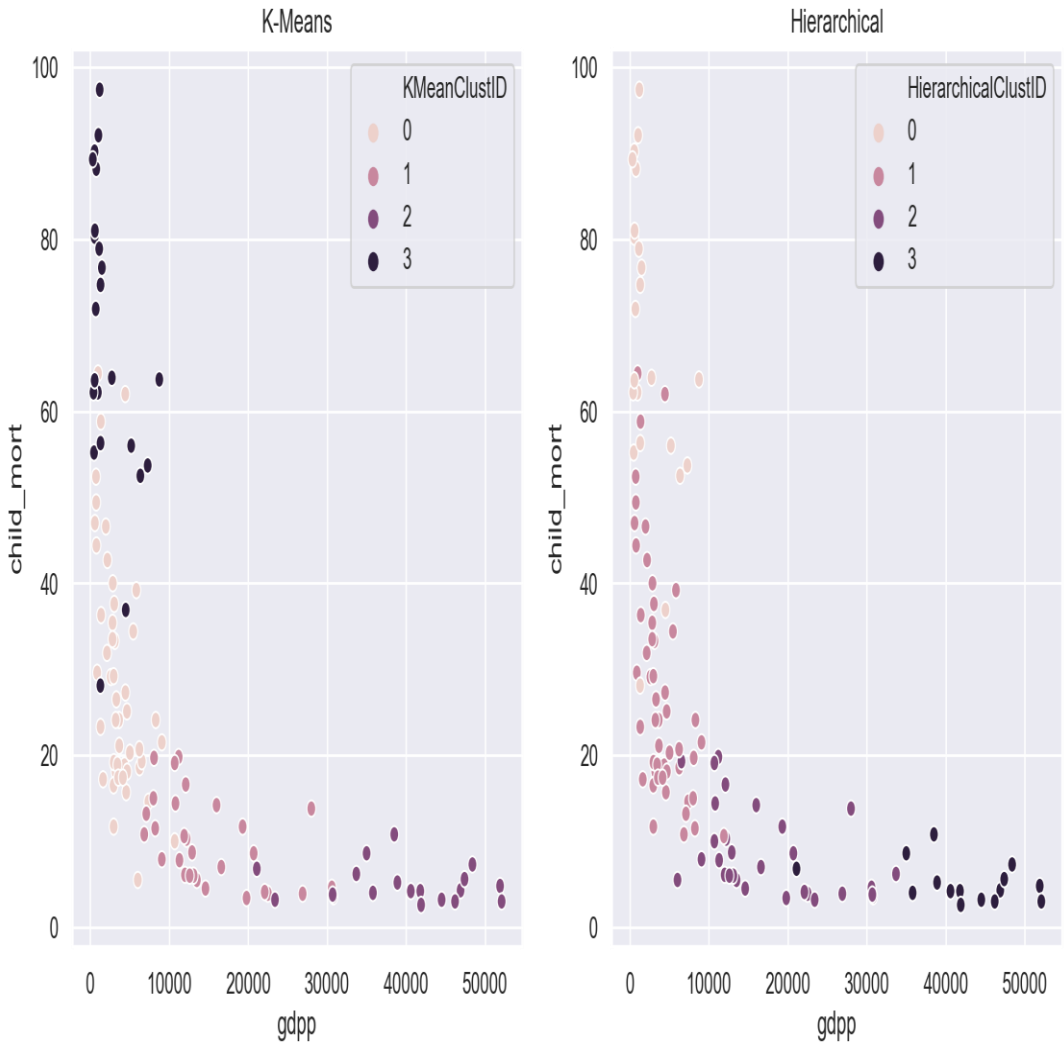
- There is not much difference in the clusters that is formed.





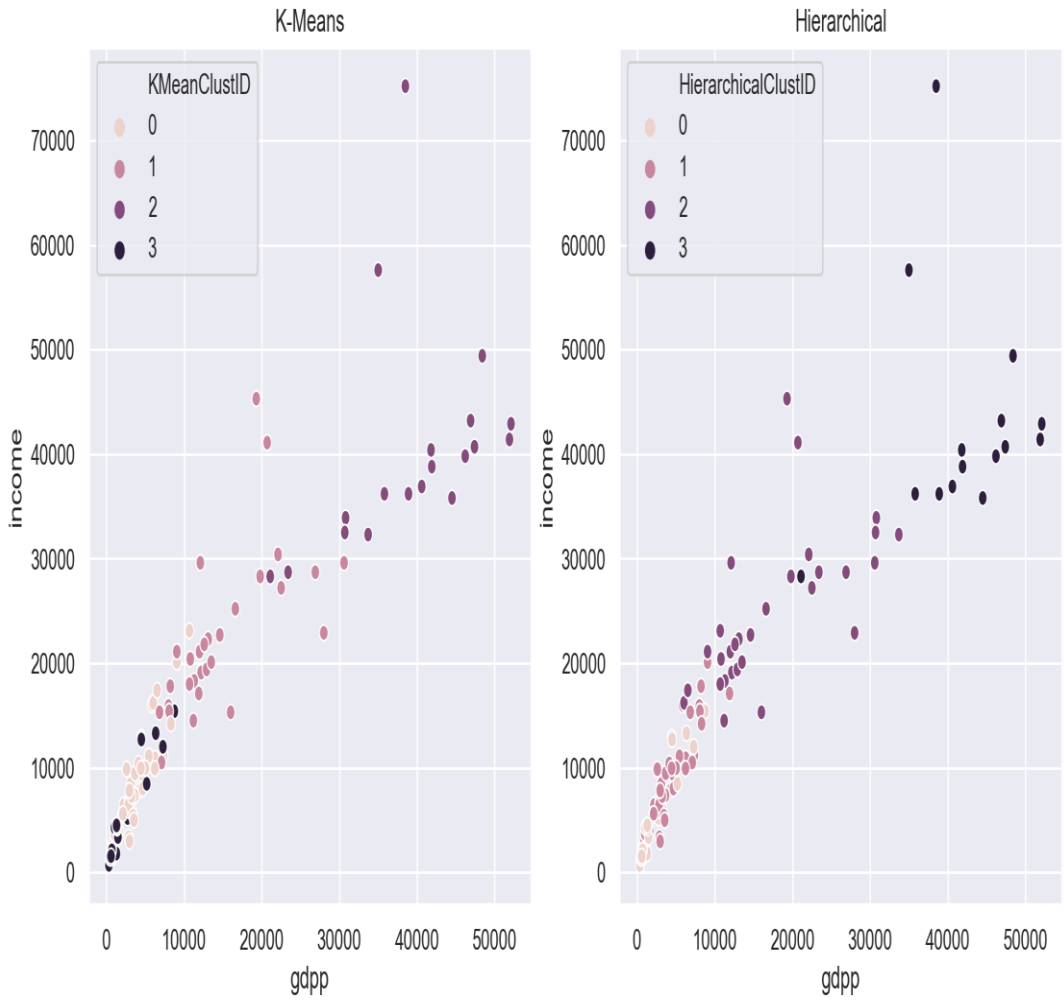
GDP Per Capita V/S Child Mortality

- Low gdpp corresponds to low household income and hence higher child mortality rate.



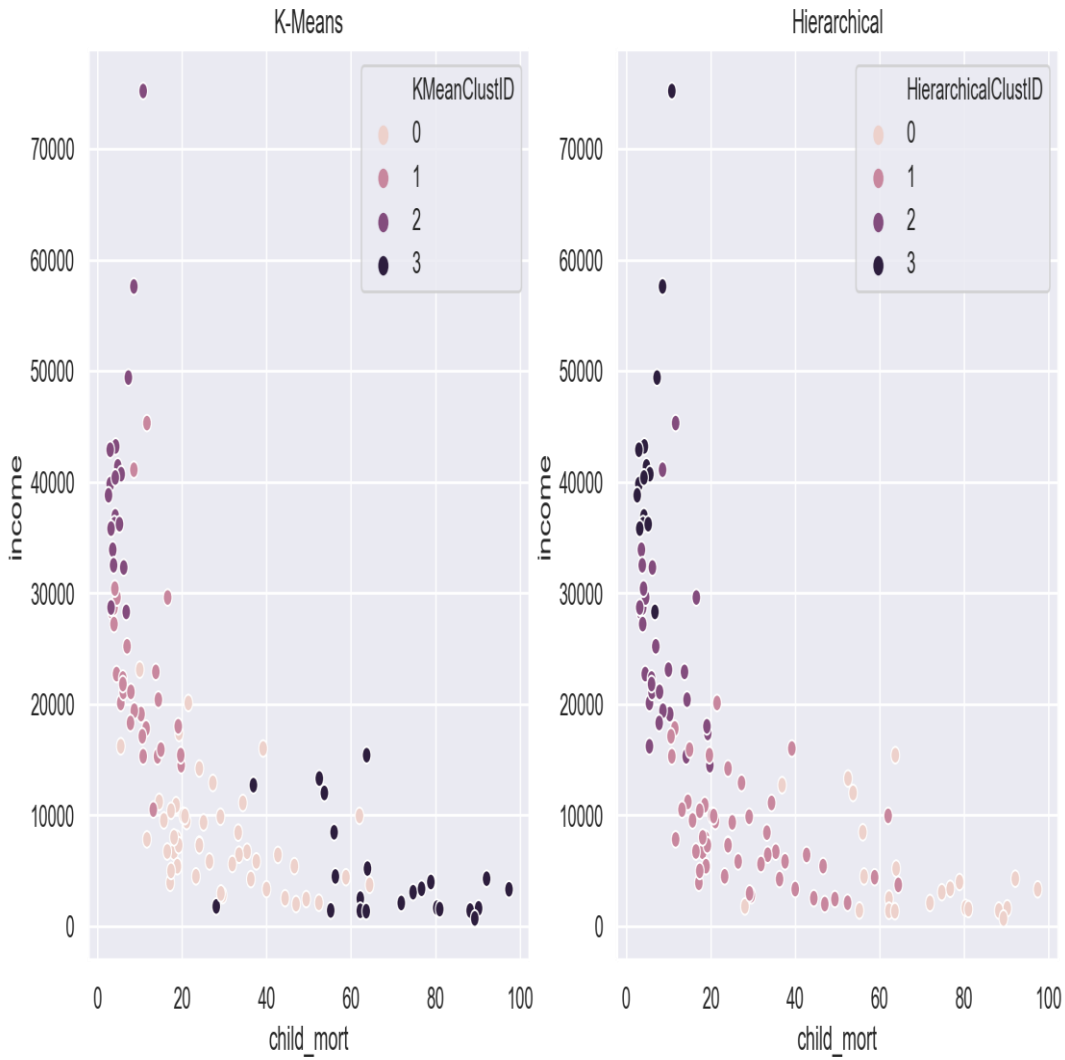
Income V/S GDP Per Capita

- We can observe a linear relationship between gdpp and income



Income V/S Child Mortality

- As we can observe from the above figure low income results in higher child mortality.
- We have analyzed both K-means and Hierarchial clustering and found clusters formed are not identical.
- The clusters formed in both the cases are not that great but its better in K-means as compared to Hierarchial.
- So, we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid.



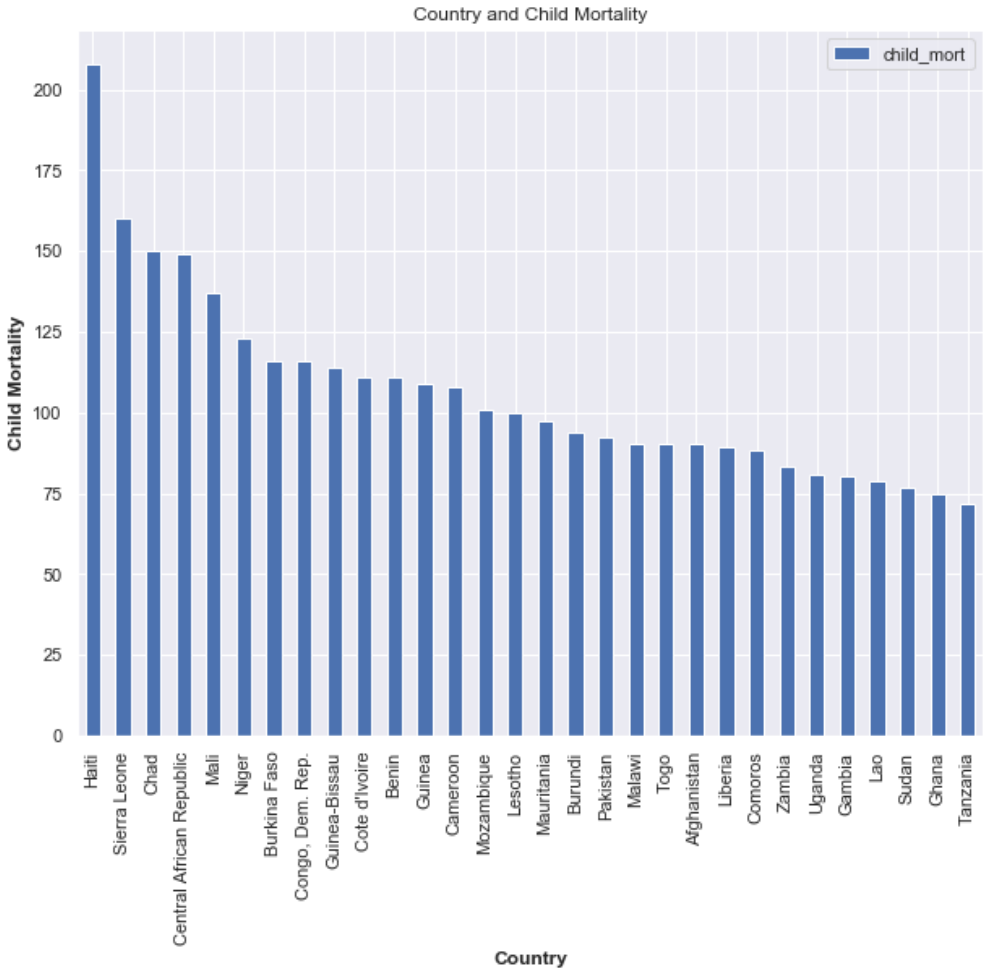
List Of Countries That Are in Need of Aid Based On Clustering

- As per both K-Means and Hierarchical Clustering common list of countries that are need of aid are as shown in pic.
- These countries have higher child mortality , low income and low GDP per capita.

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	HierarchicalClustID	KMeanClustID
0	Afghanistan	90.2000	55.3000	248.2970	41.9174	1610	9.4400	56.2000	5.8200	553	0	3
1	Botswana	52.5000	2768.6000	3257.5500	527.0500	13300	8.9200	57.1000	2.8800	6350	0	3
2	Comoros	88.2000	126.8850	397.5730	34.6819	1410	3.8700	65.9000	4.7500	769	0	3
3	Congo, Rep.	63.9000	2331.7400	1498.7800	67.4040	5190	20.7000	60.4000	4.9500	2740	0	3
4	Eritrea	55.2000	23.0878	112.3060	12.8212	1420	11.6000	61.7000	4.6100	482	0	3
5	Gabon	63.7000	5048.7500	1653.7500	306.2500	15400	16.6000	62.9000	4.0800	8750	0	3
6	Gambia	80.3000	133.7560	239.9740	31.9778	1660	4.3000	65.5000	5.7100	562	0	3
7	Ghana	74.7000	386.4500	601.2900	68.3820	3060	16.6000	62.2000	4.2700	1310	0	3
8	Iraq	36.9000	1773.0000	1534.5000	378.4500	12700	16.6000	67.2000	4.5600	4500	0	3
9	Kenya	62.2000	200.1690	324.9120	45.9325	2480	2.0900	62.8000	4.3700	967	0	3
10	Lao	78.9000	403.5600	562.0200	50.9580	3980	9.2000	63.8000	3.1500	1140	0	3
11	Liberia	89.3000	62.4570	302.8020	38.5860	700	5.4700	60.8000	5.0200	327	0	3
12	Madagascar	62.2000	103.2500	177.5900	15.5701	1390	8.7900	60.8000	4.6000	413	0	3
13	Mauritania	97.4000	608.4000	734.4000	52.9200	3320	18.9000	68.2000	4.9800	1200	0	3
14	Namibia	56.0000	2480.8200	3150.3300	351.8820	8460	3.5600	58.6000	3.6000	5190	0	3
15	Pakistan	92.1000	140.4000	201.7600	22.8800	4280	10.9000	65.3000	3.8500	1040	0	3
16	Rwanda	63.6000	67.5600	168.9000	59.1150	1350	2.6100	64.6000	4.5100	563	0	3
17	Solomon Islands	28.1000	635.9700	1047.4800	110.2950	1780	6.8100	61.7000	4.2400	1290	0	3
18	South Africa	53.7000	2082.0800	1994.7200	650.8320	12000	6.3500	54.3000	2.5900	7280	0	3
19	Sudan	76.7000	291.5600	254.5600	93.5360	3370	19.6000	66.3000	4.8800	1480	0	3
20	Tanzania	71.9000	131.2740	204.2820	42.1902	2090	9.2500	59.3000	5.4300	702	0	3
21	Uganda	81.0000	101.7450	170.1700	53.6095	1540	10.6000	56.8000	6.1500	595	0	3
22	Yemen	56.3000	393.0000	450.6400	67.8580	4480	23.6000	67.5000	4.6700	1310	0	3

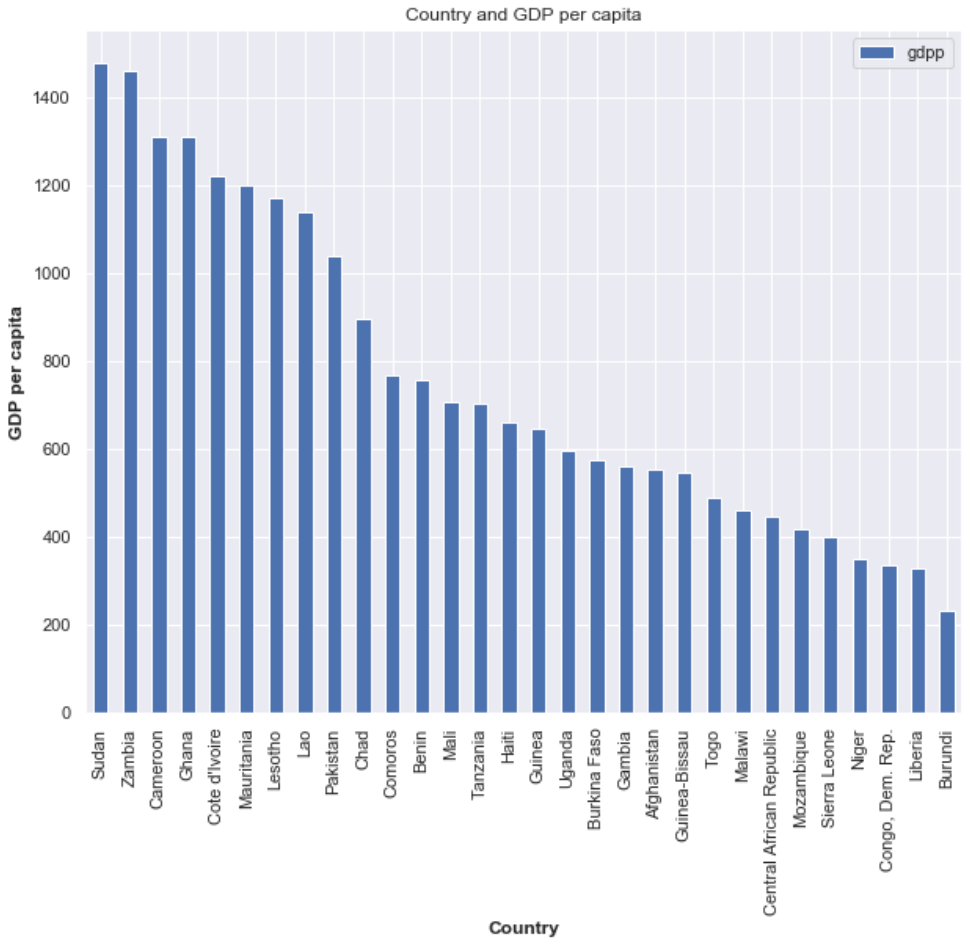
Country V/S Child Mortality

- Haiti has highest child mortality around 200+.
- Second is Sierra Leone , then comes Chad and Central African republic.
- Least is Solomon Islands.



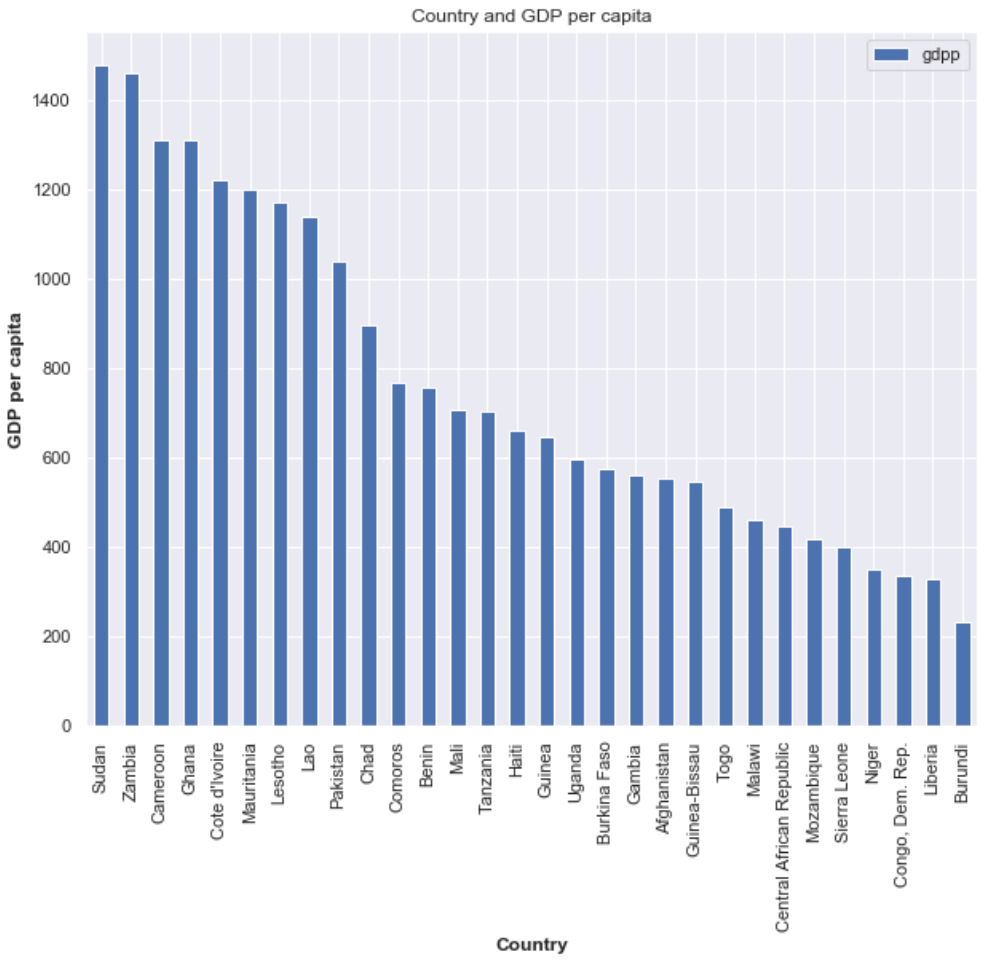
Country V/S per capita Income

- Pakistan has highest per capita income.
- Lao, Sudan comes later.
- Congo,Dem.Rep. , Liberia , Burundi , Niger , Cental african Republic has lowest income.



Country V/S GDP per capita

- Sudan has highest GDP per capita.
- Zambia, Cameroon comes later.
- Liberia , Burundi , Niger , Central African Republic. have least GDP.



Final List Of Countries Based On Socio-Economic Factors

- We have used PCA above to reduce the variables involved and then done the clustering of countries based on those Principal components and then later we identified few factors like child mortality, income etc which plays a vital role in deciding the development status of the country and builded clusters of countries based on that.
- Based on those clusters we have identified the below list of countries which are in dire need of aid.
- The list of countries are subject to change as it is based on the few factors like Number of components chosen, Number of Clusters chosen, Clustering method used etc.which we have used to build the model.

```
0      Afghanistan
1      Benin
2      Burkina Faso
3      Burundi
4      Cameroon
5      Central African Republic
6      Chad
7      Comoros
8      Congo, Dem. Rep.
9      Cote d'Ivoire
10     Gambia
11     Ghana
12     Guinea
13     Guinea-Bissau
14     Haiti
15     Lao
16     Lesotho
17     Liberia
18     Malawi
19     Mali
20     Mauritania
21     Mozambique
22     Niger
23     Pakistan
24     Sierra Leone
25     Sudan
26     Tanzania
27     Togo
28     Uganda
29     Zambia
Name: country, dtype: object
```